We claim:

1. In a data processing system wherein descriptor vectors associated with a plurality of regions of molecules are stored in a database, a method for generating and storing data characterizing at least one region of said plurality of regions, the method comprising the steps of:

generating an entry comprising i) an identifier that identifies said at least one region, and ii) data characterizing a set of axes derived from property distribution of said at least one region;

applying a mapping the descriptor vector associated with said at least one region;

generating a key that corresponds to said mapping of the descriptor vector associated with said at least one region; and

storing said entry in a memory, wherein said key is associated with said entry.

2. The method of claim 1, wherein said set of axes are invariant to rotation and translation of said at least one region.

3. The method of claim 2, wherein said set of axes are derived from principal axes of said property distribution.

4. The method of claim 3, wherein said property distribution of said at least one region is based upon application of a smearing function to a property field.

5. The method of claim 1, wherein said plurality of descriptor vectors are classified into groups, wherein said mapping step maps said descriptor vector to a said space optimally discriminates between said groups of descriptor vectors.

6. The method of claim 5, wherein said mapping is derived from the steps of:

generating first data representing differences between said groups of descriptor vectors;

generating second data representing variations within said groups of descriptor vectors;

identifying a set of component vectors that maximizes an F distributed criterion function, said criterion function having a numerator based upon said first data and a denominator based upon said second data;

generating an F distributed statistic for subsets of said component vectors, said statistic having a numerator based upon said first data and a denominator based upon said second data;

for each particular subset of component vectors, calculating a probability value for the F-distributed statistic associated with the particular subset;

selecting a probability value from probability values for said subsets of component vectors based upon a predetermined criterion;

identifying the subset of said component vectors associated with the selected probability value; and

generating a mapping to a space corresponding to the subset of component vectors associated with the selected probability value, and storing the mapping for subsequent processing.

7. The method of claim 6, wherein said first data comprises a matrix ε_b representing covariance between said groups of descriptor vectors, and said second data comprises a matrix ε_w representing covariance within said groups of descriptor vectors.

8. The method of claim 7, wherein said criterion function has the general form:

$$f(\hat{w}) = C\left(\frac{\hat{w}^T \varepsilon_b \hat{w}}{\hat{w}^T \varepsilon_w \hat{w}}\right)$$

where $\hat{w}$ is some vector, and C is a constant based upon degrees of freedom in ε_b and ε_w.

9. The method of claim 8, wherein C is determined as follows:

$$C = \frac{1/\text{degrees of freedom in } \varepsilon_b}{1/\text{degrees of freedom in } \varepsilon_w} = \frac{1/(N-1)}{1/(\sum n_i - N)}$$

where N represents the number of groups of descriptor vectors, n_i represents the number of regions, and $\sum n_i$ represents the sum of n_i for the N groups.

10. The method of claim 7, wherein the step of identifying a set of component vectors that maximizes an F distributed criterion function comprises the substeps of:

determining a set of (eigenvalue, eigenvector) pairs for the matrix ε_w

determining said set of component vectors based upon said set of (eigenvalue, eigenvector) pairs for the matrix ε_w.

11. The method of claim 10, wherein said statistic for a given subset of component vectors is based upon value of said criterion function for said subset of component vectors.

12. The method of claim 11, wherein said statistic for a given subset of component vectors has the following form:

$$\psi_S = C\left(\frac{1}{L_S}\right) \sum f_k$$

where f_k represents the value of the criterion function at a component vector in the given subset,

C is a constant,

L_S represents the number of f_k values in the given subset of component vectors, and

the $\sum$ operation sums over the L_S f_k values in the given subset of component vectors.

13. The method of claim 12, wherein said a probability value for a particular F-distributed statistic represents a probability value that the particular F-distributed statistic could have been larger by chance.

14. The method of claim 13, wherein said probability value selected from probability values for said subsets of component vectors is a minimum probability value of said probability values for said subsets of component vectors.

15. The method of claim 6,

wherein said mapping for said at least one descriptor vector performs a loop over each component vector belonging to the subset of component vectors associated with the selected probability;

wherein, in each iteration of said loop, dot product of said descriptor vector with a transpose of a unit vector for the given component vector is added to a running sum.

16. In a data processing system wherein descriptor vectors associated with a plurality of regions of molecules are stored in a database, CHARACTERIZED IN THAT said data processing system includes a memory storing a plurality of entries each comprising i) an identifier that identifies at least one region and ii) data characterizing a set of axes derived from property distribution of said at least one region, a method for determining alignment of similar molecular structure comprising the steps of:

providing a descriptor vector associated with said query molecular region;

mapping said descriptor vector associated with said query molecular region;

generating a second key that corresponds to said mapping of said descriptor vector associated with said query molecular region; and

retrieving from said memory entries that are associated with a first key that corresponds to said second key; and

for at least one entry retrieved from said memory,

generating data that represents a match hypothesis associated with said query molecular region and at least one region R identified by said at least one entry retrieved from said memory, wherein said data is based upon parameters of a transformation that aligns a set of axes derived from property distribution of said query molecular region with a set of axes derived from property distribution of said at least one region R,

determining a score associated with said data, and

storing said data and score as an entry in a vote table.

17. The method of claim 16, further comprising the step of:

selecting one or more entries of said vote table based upon said score associated with said entries; and

identifying at least one region that corresponds to the selected entries of said vote table as a potential matching regions to said query molecular region.

18. The method of claim 16, wherein said set of axes derived from property distribution of a region are invariant to rotation and translation of said region.

19. The method of claim 18, wherein said set of axes derived from property distribution of a region are derived from principal axes of said property distribution.

20. The method of claim 19, wherein said property distribution of said region is based upon application of a smearing function to a property field.

21. The method of claim 16, wherein said plurality of descriptor vectors are classified into groups, and wherein said mapping step maps said descriptor vector to a space optimally discriminates between said groups of descriptor vectors.

22. The method of claim 21, wherein said mapping is derived from the steps of:

generating first data representing differences between said groups of descriptor vectors;

generating second data representing variations within said groups of descriptor vectors;

identifying a set of component vectors that maximizes an F distributed criterion function, said criterion function having a numerator based upon said first data and a denominator based upon said second data;

generating an F distributed statistic for subsets of said component vectors, said statistic having a numerator based upon said first data and a denominator based upon said second data;

for each particular subset of component vectors, calculating a probability value for the F-distributed statistic associated with the particular subset;

selecting a probability value from probability values for said subsets of component vectors based upon a predetermined criterion;

identifying the subset of said component vectors associated with the selected probability value; and

generating a mapping to a space corresponding to the subset of component vectors associated with the selected probability value, and storing the mapping for subsequent processing.

23. The method of claim 22, wherein said first data comprises a matrix ε_b representing covariance between said groups of descriptor vectors, and said second data comprises a matrix ε_w representing covariance within said groups of descriptor vectors.

24. The method of claim 23, wherein said criterion function has the general form:

$$f(\hat{w}) = C\left(\frac{\hat{w}^T \varepsilon_b \hat{w}}{\hat{w}^T \varepsilon_w \hat{w}}\right)$$

where $\hat{w}$ is some vector, and C is a constant based upon degrees of freedom in ε_b and ε_w.

25. The method of claim 24, wherein C is determined as follows:

$$C = \frac{1/\text{degrees of freedom in } \varepsilon_b}{1/\text{degrees of freedom in } \varepsilon_w} = \frac{1/(N-1)}{1/(\sum n_i - N)}$$

where N represents the number of groups of descriptor vectors, n_i represents the number of regions, and $\sum n_i$ represents the sum of n_i for the N groups.

26. The method of claim 23, wherein the step of identifying a set of component vectors that maximizes an F distributed criterion function comprises the substeps of:

determining a set of (eigenvalue, eigenvector) pairs for the matrix ε_w

determining said set of component vectors based upon said set of (eigenvalue, eigenvector) pairs for the matrix ε_w .

27. The method of claim 26, wherein said statistic for a given subset of component vectors is based upon value of said criterion function for said subset of component vectors.

28. The method of claim 27, wherein said statistic for a given subset of component vectors has the following form:

$$\psi_S = C\left(\frac{1}{L_S}\right) \sum f_k$$

where f_k represents the value of the criterion function at a component vector in the given subset,
C is a constant,
L_S represents the number of f_k values in the given subset of component vectors, and
the $\sum$ operation sums over the L_S f_k values in the given subset of component vectors.

28. The method of claim 22, wherein said a probability value for a particular F-distributed statistic represents a probability value that the particular F-distributed statistic could have been larger by chance.

29. The method of claim 28, wherein said probability value selected from probability values for said subsets of component vectors is a minimum probability value of said probability values for said subsets of component vectors.

30. The method of claim 22,

wherein said mapping for said at least one descriptor vector performs a loop over each component vector belonging to the subset of component vectors associated with the selected probability;

wherein, in each iteration of said loop, dot product of said descriptor vector with a transpose of a unit vector for the given component vector is added to a running sum.